The Horizontal Pod Autoscaler automatically scales the number of pods in a replication controller, deployment or replica set based on observed CPU utilization (or, with custom metrics support, on some other application-provided metrics).

**Deploy the metrics server**

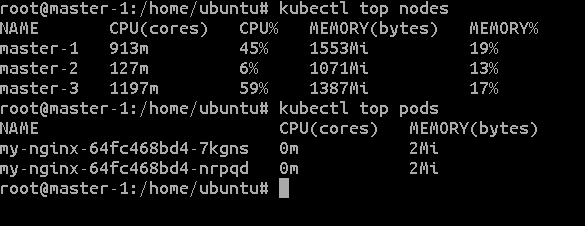
Replace **<your-name>** with your **name** throughout the Lab.

1. SSH to the **haproxyVM on Azure or any of the master NODE** of the 3 AWS Clusters

|  |
| --- |
| $ sudo su # git clone https://github.com/LovesCloud/metrics-server.git  # kubectl create -f metrics-server/metrics-server/ # kubectl get po -n kube-system |

**2. Wait for 5 Mins and run the below commands**

|  |
| --- |
| # kubectl top nodes |



**3. Run & expose the Application**

Create a new deployment with the below command and replace <your-deployment-name> with

unique deployment name.

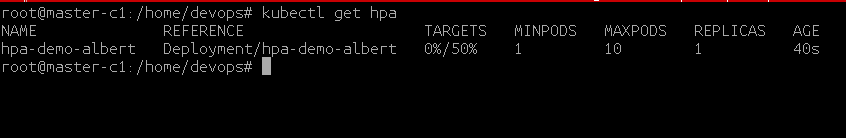
|  |
| --- |
| # kubectl run hpa-demo-<your-name> --image=k8s.gcr.io/hpa-example --requests=cpu=200m --port=80  # kubectl expose deploy hpa-demo-<your-name> --type=NodePort |

## **4. Create Horizontal Pod Autoscaler**

|  |
| --- |
| # kubectl autoscale deployment hpa-demo-<your-name> --cpu-percent=50 --min=1 --max=10 |

5. Please wait for 2-3 minutes before running the below command

|  |
| --- |
| **# kubectl get hpa** |

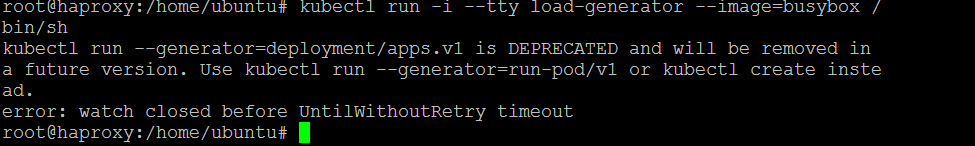


Please note that the current CPU consumption is 0% as we are not sending any requests to the server (the CURRENT column shows the average across all the pods controlled by the corresponding deployment).

**6. Increase load**

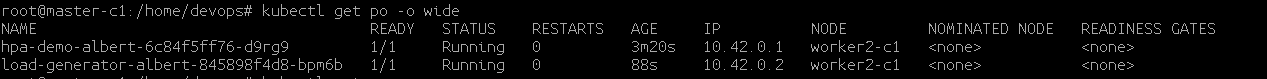
|  |
| --- |
| **# kubectl run -i --tty load-generator**-<your-name> **--image=busybox /bin/sh** |

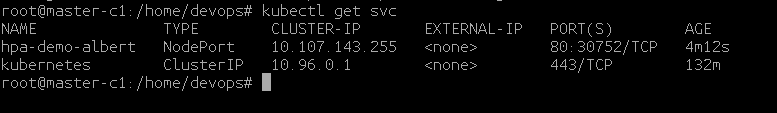
7. IF you get the below error follow stop 9 and 10 to reattach the load generator, else continue from below.



**8. SSH to master/haproxy VM on a different terminal and check your NODE’s Internal-IP and NodePort of the HPA deployment run the below command** ( example worker1’s internal IP in this example)

|  |
| --- |
| # kubectl get po -o wide  # kubectl get svc |





**Write down the below command on a NotePad for the while loop replace NODE-internal-ip and NodePort obtained from the above commands**.

|  |
| --- |
| while true; do wget -q -O- http://<NODE-internal-ip>:<NodePort>/ ; done |

**Where, NODE-internal-ip is the internal IP of the Node where the hpa-demo PODs are deployed NodePort is the Port on which the hpa-demo deployment is exposed to, in this Example worker2-c1’s internal IP and 32121 is the NodePort**

Example

|  |
| --- |
| while true; do wget -q -O- http://172.31.69.149:30752/ ; done |

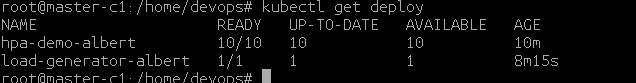
**9. Reattach the load generator**

Run the below command to get the load-generator pod name

|  |
| --- |
| **# kubectl get pods** |



|  |
| --- |
| # kubectl get deploy |



Where **load-generator-albert-845898f4d8-bpm6b**  is the pod name and **load-generator (default)** is the load generator deployment name in this example.

10. Update only the “**load-generator-557649ddcd-vk5m6**” and <your-name> with your POD name and your name respectively in the below command.

|  |
| --- |
| **kubectl attach load-generator-557649ddcd-vk5m6 -c load-generator-<your-name> -t -i** |

11. On the **Load generator POD CLI** past the command saved in step-2 above to increase load again.

|  |
| --- |
| # while true; do wget -q -O- http://<NODE-internal-ip>:<NodePort>/ ; done |

Example

|  |
| --- |
| # while true; do wget -q -O- http://172.31.69.149:30752/ ; done |

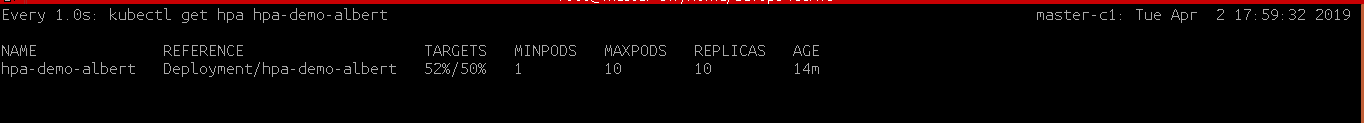


**Do not exit the load generator.**

**12. Launch** **two more terminals** of the HAProxy VM/master instance and run the below commands  
Within a minute or so, we should see the higher CPU load by executing:

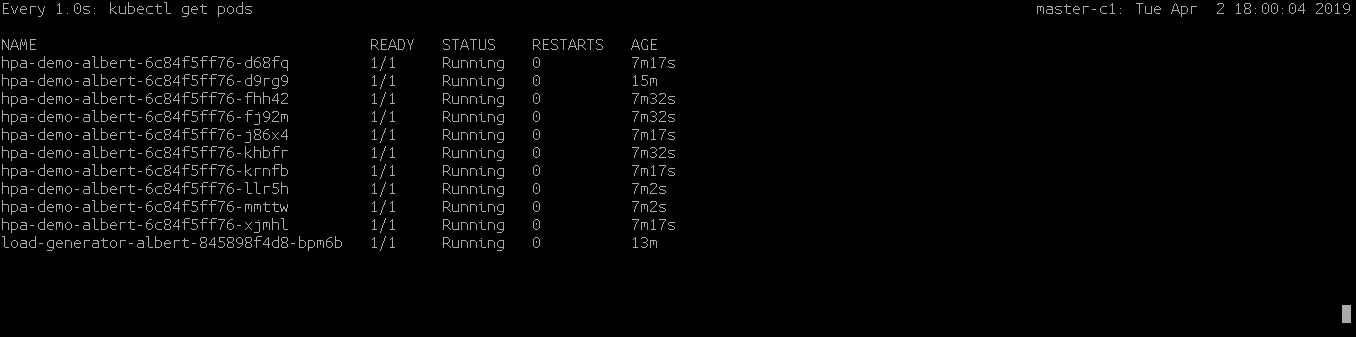
13. On terminal one run the below command

|  |
| --- |
| $ sudo su # watch -n 1 kubectl get hpa hpa-demo-albert |



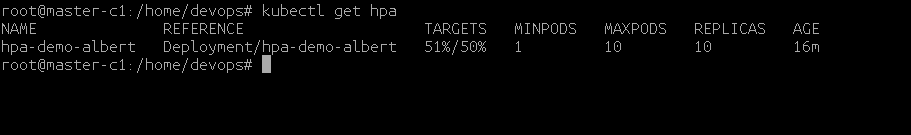
**14**. On terminal two run the below command

|  |
| --- |
| $ sudo su  # watch -n 1 kubectl get pods |



15. Press Ctrl+c on one of the two terminals from above and run the below commands.

|  |
| --- |
| # kubectl get hpa |



16. Get the deployment details.

|  |
| --- |
| # kubectl get deployment hpa-demo-<your-name> |

****

**18.** Login to the Kubernetes Dashboard and you will be able to observe that the application has started scaling horizontally.

**19.** Delete the Deployments and HPA

|  |
| --- |
| # kubectl delete deploy hpa-demo-<your-name> # kubectl delete hpa hpa-demo-<your-name> |